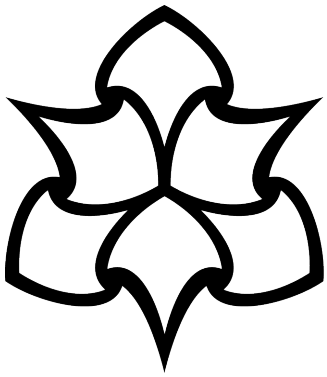


Trust in Computational Intelligence Systems: A Case Study in Public Perceptions

Keeley Crockett[1], Sean Goltz[2], Matt Garratt[3], Annabel Latham[1]

[1]School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, M1 5GD, UK, K.Crockett@mmu.ac.uk
[2]Business & Law School, Edith Cowan University, Perth, Australia, n.goltz@gmail.com
[3]School of Engineering and IT, University of New South Wales, PO Box 7916, Canberra BC 2610, ACT 2902, Australia, m.garratt@adfa.edu.au

***Abstract*— The public debate and discussion about trust in Computational Intelligence (CI) systems is not new, but a topic that has seen a recent rise. This is mainly due to the explosion of technological innovations that have been brought to the attention of the public, from lab to reality usually through media reporting. This growth in the public attention was further compounded by the 2018 GDPR legislation and new laws regarding the right to explainable systems, such as the use of "*accurate data*", "*clear logic*" and the "*use of appropriate mathematical and statistical procedures for profiling*". Therefore, trust is not just a topic for debate – it must be addressed from the onset, through the selection of fundamental machine learning processes that are used to create models embedded within autonomous decision-making systems, to the selection of training, validation and testing data. This paper presents current work on trust in the field of Computational Intelligence systems and discusses the legal framework we should ascribe to trust in CI systems. A case study examining current public perceptions of recent CI inspired technologies which took part at a national science festival is presented with some surprising results. Finally, we look at current research underway that is aiming to increase trust in Computational Intelligent systems and we identify a clear educational gap.**

Keywords- Ethics, Trust, Explainability, Morality, Computational Intelligence, GDPR

I. INTRODUCTION

During the past year, there has been an increase in the attempts to understand, define, and analyse what constitutes trust in artificial intelligence (AI) algorithms, among industry, academia, the media and the public (we use CI and AI as identical for the purpose of this article. See explanation below). The consensus is that the future of AI holds benefits to humans that may be larger than the expected harm. This positive balance can be achieved if the human race would ensure transparency, safety, privacy, as well as remove or mitigate bias and take ethical considerations seriously [1]. Tschopp writes *"Although there may be clear benefits for humanity, like defeating cancer or halting climate change, AI is often viewed with great skepticism as the hype around AI leads to justified resistance*" [1]. In this contested realm, trust is the key to shift the balance between advantages and disadvantages. If AI based systems are to succeed in areas that can benefit human quality of life, they must be trusted.

In this paper, we recognize that the public are generally less familiar with the concept of computational intelligence, but are much more familiar with the term artificial intelligence (AI). Traditionally CI was supposed to be about 'soft computing' and AI was about 'hard computing,' but the lines are often blurred as many of the underlying algorithms are the same. Jim Bezdek who is credited with one of the first clear definitions of CI in 1994 stated that "*I think the debate about 'CI vs. AI', including questions such as 'Is it CI or AI?'* or '*Does one of these areas include the other, do they overlap, etc.?' are really pretty moot nowadays*" [2]. Hence, in this paper we assume that the underlying issues of trust are the same for both taxonomies and the survey questions discussed later use the acronym AI rather than CI.

For the purposes of our discussion, Trust in CI systems refers to the trust that a human being places in a CI system when interacting with it. In the more general case, trust can also refer to trust between CI agents or trust of a human by a CI agent. Humans are capable of both over-trust and under-trust of CI systems based on many factors including exposure, training and human bias. Both cases can be dangerous. For example, when a human over-trusts a self-driving car autopilot system, they can fail to provide necessary oversight leading to an accident. Conversely, a human that ignores an automated safety warning system because of lack of trust, can lead to accidents. Lack of trust can also slow the take-up of technology or limit its use which can create a missed opportunity in terms of the potential financial and social benefits of technology. Rousseau et al. defines trust as, *"a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another"* [3]. Luhmann contends that trust should be understood in connection with the human attempt to adapt in an effective way to the environment by reducing complexity and uncertainty [4], while Gambetta argues that trust concerns *"the subjective probability with which an agent assesses that another agent or group of agents will perform a particular*

action" [5]. For the purpose of this paper, we adopt Gambetta's 'simple' definition as most of the survey questions refer to the probability of the positive performance of a particular action. For example, whether an autonomous vehicle will drive me safely to the requested destination.

The concept of fairness in terms of developing CI systems must also be considered in this discussion [6]. Corbett-Davie et al. [7] describes the three predominant categories of fairness: anti-classification – associated with risk assessment algorithms not including personal sensitive data, such as gender, when making calculations. This data is known as protected attributes; classification parity – which requires that "*certain common measures of predictive performance be equal across group*s" [7] and finally calibration – which states that outcomes are independent across protected attributes – which implies in principle that a system making an automated decision should be the same for all personal sensitive data including ethnicity, gender etc.

This paper first examines current CI system applications that have been covered by academics, industry and the media where levels of trust have been debated (Section II). The legalities of the concepts of trust are overviewed to emphasize the complexities of human - CI Trust and the regulatory mechanisms and challenges associated with this realm across three main aspects: 1) the legal status of the CI agent; 2) transparency and human rights; 3) legal accountability for harm caused by CI.

Public perceptions of trust are an important factor in how we communicate CI applications to the average person who is likely to have no background in the field and whose opinion is often driven by how the media report the application. Indeed, how a news story is interpreted and reported by a journalist could be factually incorrect, which could lead to a waterfall effect of fake news which in turn drives public belief. In Section III, we present the results of a 2018 case study held at a National Science Festival in the UK, which looks at the general public perceptions of Trust in CI systems. The study comprises of a survey which was designed to reveal what level of trust the public has in CI systems. The study presented to members of the public twenty questions each relating to the latest state-of-the-art developments in CI and asked them which ones they trust and was designed to stimulate conversation and debate. Each question had a related image captured through an associated news story that reported factually, a CI application. Participation was voluntary, where completely random members of the public opted in to take part and no incentives were given. Finally, in Section IV we look at ways the CI community can increase public awareness of, and trust in, CI systems.

II. TRUST IN CI SYSTEMS

A. CI System applications

The applications of CI systems are numerous and rapidly expanding with great potential for reducing human workload and eliminating mundane tasks. Applications include assessing the probability of a prisoner to re-offend, predicting stock prices, marking essays, conducting medical diagnosis and control of autonomous vehicles. The consequences of an error by CI system can range from mildly inconvenient (e.g., unfairly rejecting a bank loan) to catastrophic (e.g., crash of a self-driving car or failure to diagnose a malignant cancer). Hence, sufficient critical human oversight of CI is needed to check that systems are operating safely and fairly. The question of trust then arises with regard to whether people's over-trust in CI can lead to errors that can be overlooked.

A number of studies have shown that people tend to over-trust robots, e.g. in [8, 9], even when they have been seen to previously fail. Conversely, in emergency situations humans can quickly become suspicious of robots when they are seen to make a minor mistake. This can lead humans to ignore the robot's advice even when the robot is genuinely and accurately trying to help them - e.g., guide them out of a burning building.

In March 2018, the first pedestrian to be killed by a self-driving car was run over by an Uber self-driving car during testing. At the time of the incident, the safety driver was looking down in what is clearly a case of over-trust in the autonomous system. A few other types of accident have recently occurred when human drivers did not pay attention when self-driving autopilots were engaged. These tasks are particularly challenging due to the difficulty for the human driver to maintain vigilance with the almost redundant task of overseeing the autonomous system over long periods. Because the task ends up being rather boring, once initial anxiety about handing over control to a computer wanes, the human driver quickly shifts to over-trust and neglect when overseeing the machine.

Chat-bots are used increasingly for applications such as health care where they can be utilized to augment the advice given by doctors, especially for follow-up counseling after a medical procedure [10]. The use of chat-bots can reduce the pressure on humans in the workforce and drive down costs. However, building trust in the chat-bot is critical towards getting the user to engage and not become frustrated and generally feel ill-will towards the service provider. If a user does not trust the chat-bot, it may turn her away from the service and seek alternative service. This, in turn, could result in financial implications for both the employer and the company offering the chat-bot. Hence, chat-bot designers deliberately build in traits that engender trust from users such as predictability, natural language traits and anthropomorphism. In the latter case, the amount of anthropomorphism is critical as too much similarity to humans can actually create an 'uncanny valley' situation where the user becomes put-off by the small differences between the chat-bot and a real human [11]. Studies have shown that people interacting with chat-bots are likely to become overly trusting, handing over personal information and even passwords without proper consideration [12]. A study at Coventry University found that students were more likely to answer survey questions about drugs or sex when talking to a chat-bot than when using a standard online questionnaire [13]. The trust relationship that people build

with chat-bots is open to exploitation by anyone wanting to acquire personal information about people whether for criminal use (e.g., identity theft) or just for targeted marketing.

B. The Legalities of Trust in CI Systems

There are three interconnected legal aspects to consider in the context of trust between a human and its interaction with CI systems: 1) the legal status of the CI agent; 2) transparency and human rights; 3) legal accountability for harm caused by CI [14]. With respect to the legal status, IEEE recommends that "*it would be unwise to accord personhood status to AI at this time. AI should therefore remain to be subject to the applicable regimes of property law."* As for transparency and human rights, it is suggested that AI should be designed to be transparent and accountable as primary objectives.

One controversial example is the software that helps judges make decisions on parole. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a 4th generation risk assessment software is being used by judges across the USA to generate risk scores of how likely an offender will re-offend. It was developed to assess crucial static and dynamic risk and needs factors and to provide support in decisions around placement, supervision, and case management [15]. The lack of transparency and alleged gender bias of COMPAS was discussed by the Court in the case of Loomis v Wisconsin. Loomis was arrested in 2013 for his involvement in a drive-by shooting and the judge sentenced him to seven years, saying he was "high risk". The judge based this analysis, in part, on the risk assessment score given by COMPAS. Loomis argued that COMPAS violated his right to due process because the proprietary nature of the COMPAS algorithm made it impossible to test its scientific validity and because the tool improperly considers gender in determining risk. The Wisconsin Supreme Court affirmed the lower court's decision that the risk assessment may be considered as one factor among many used in sentencing [16]. The unanimous court also concluded that the tool did not violate Loomis' due process right to not be sentenced on the basis of gender. The court wrote of COMPAS: *"If the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose"*. Finally, the US Supreme Court declined to hear Loomis' appeal. It seems that the Court's decision was influenced by the Solicitor-General amicus curiae brief arguing that given *"the highly limited purpose for which petitioner's ability to counter the factual information on which the assessment relied, the Wisconsin Supreme Court correctly declined to find a due process violation. But that is not to say that the use of actuarial risk assessments at sentencing will always be constitutionally sound*" [17].

The third aspect, legal accountability to harm, was traditionally dealt between humans based on two principles: 1) Alterum non Laedere ("do not injure another") and 2) mechanisms of the burden of proof. These principles have been incorporated to distribute risk and responsibility. Applying these principles to CI systems is challenging due to their nature and raises questions of liability of third parties (manufacturer, operator etc). Current discussions into the legalities of CI system in the context of trust and interactions with humans, raises additional questions of privacy as well as ethical questions regarding sex robots, for example [18]. Furthermore, recent research found that 20% of consumers would either definitely, or possibly, trust a CI system to provide advice on a legal case relating to them [19]. According to Pagallo, in the agent's attempt to adapt to the environment, the trustor 'delegates' some actions that are necessary for achieving a specific goal or gain [20]. This sort of trust delegation should not be interpreted in strict legal terms [21]. Furthermore, Pagallo contends that, "*both the unpredictability of robots' behavior and their capability to act on human behalf call for a rethinking of the traditional legal framework*" [20].

III. CASE STUDY: PUBLIC PERCEPTION ON TRUST

A. Overview

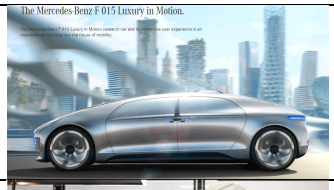
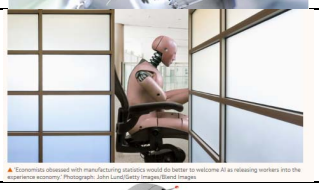
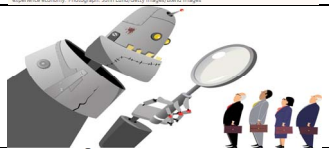
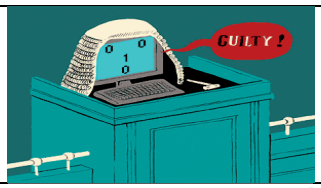
The Manchester Science festival in the UK is an annual event attracting around 130,000 visitors each year. This year Manchester Metropolitan University won a bid to stage a Platform of Investigation entitled "Me Verses the Machine" on Saturday 20th October (an all-day event) at the Science and Industry Museum, Manchester, UK. The bid included eight STEM activities designed to introduce families to artificial intelligence, coding and computer science through offering hands on activities. One activity was the Great Computational Intelligence debate where we invited members of the public to engage with researchers and academics to discuss and debate ten topical questions on recent applications of CI technology. This activity comprised an HDTV to present the questions with images, a table of further hard copy questions and images to make the study accessible and anonymous question sheets to record their opinions. The aim was to seek the opinions of members of the public towards recent applications of Computational Intelligence in terms of which situations they felt they would be able to trust the technology.

B. Methodology

Twenty questions were formulated covering a wide area of society where new CI innovations had been reported in the media such as self-driving vehicles, healthcare, policing, teaching, space exploration, warfare etc. These questions, shown in Table I, all included an image associated with an article, which described a new technology or innovation involving computational intelligence. In Table I, we have also included article references to justify our selection of questions within the survey. The members of the public did not read the article before answering the question but there was an opportunity to take away a copy of the questions with links to the articles if they were interested. Participation by members of the public was voluntary. To take part they approached the exhibit, had the process explained and then decided to take part. During the debate, the public were asked to record anonymously the answers to 20 questions (Table I) which could be "yes" – indicating they would trust the CI

system in the situation, “no” indicating they did not trust the system, or “Abstain” meaning they weren’t sure, did not have an opinion or they felt there was currently insufficient evidence to make a decision.

Table I: 20 questions

Q-No	Question / Source	Image
1	Would you let your Grandma use a Self-Driving Wheelchair ? Source: [22]	
2	Would you like an Eldercare Robot to look after your grandparents so they could stay in their own home? Source: [23]	
3	Would you like your Pizza delivered by an Uber robot? [24]	
4	Q4. Do you think we should use Drones to Spy on people? [25]	
5	Would you like to speak to a robot administrator when you visit your GP? [26]	“ „websites and artificial intelligence “chat bots” could replace up to 90% of Whitehall’s administrators, as well as tens of thousands in the NHS and GPs’ surgeries, by 2030 – saving as much as £4bn a year.” – the Guardian
6	Would you trust a a self-driving car? [27]	
7	Would you like Robot to make your pizza? [28]	
8	Should robots be used for Space Exploration instead of humans? [29]	
9	Do you trust a Robot Surgeon to operate on you? [30]	
10	Are you worried that robots will steal your job? [31]	
11	Do you think that machines can be biased? [32]	
12	Would you send your kids to School in a self-driving bus? [33]	
13	Would you like a robot to support your health care in a hospital? [34]	
14	Should a machine be liable in a court of law ? [35]	
15	Would you like to be interviewed by a robot police officer? [36]	

16	Do you think robots can predict cancer better than one human doctor? [37]	
17	Would you like to be taught a course by a Robot tutor? [38]	
18	Would you prefer a robot pet to a real one? [39]	
19	Would you let AI choose an outfit for you for your best friend's wedding? [40]	
20	Would using facial recognition technology at your child's school make you feel your child was safer in school? [41]	

C. Results

During the six-hour event, 625 visitors (415 adults, 210 children) passed through the Platform for Investigation (PI) Exhibition. 68 members of the public volunteered to take part in the Great CI debate that was indicated through completion of the question sheets. The sample was completely random in that the researchers had no prior knowledge of members of the public whom may or may not attend the science festival, the PI exhibition, and finally who would self-volunteer to take part. It was found through initial conversation that all of the public knew what the term Artificial Intelligence was, but did not necessarily understand in detail what it meant. Table II shows the average results for the categories (Yes, No, Abstain) for the 20 questions.

Table II: Results

Q-No	Question	Yes	No	Abstain
1	Would you let your Grandma Use a Self-Driving Wheelchair?	36%	50%	14%
2	Would you like an Eldercare Robot to look after your grandparents so they could stay in their own home?	59%	32%	9%
3	Would you like your Pizza delivered by an Uber robot?	68%	27%	5%
4	Do you think we should use Drones to Spy on people?	46%	27%	27%
5	Would you like to speak to a robot administrator when you visit your doctor?	59%	32%	9%
6	Would you trust a self-driving car?	27%	0%	73%
7	Would you like Robot to make your pizza?	77%	19%	4%
8	Should robots be used for Space Exploration instead of humans?	59%	36%	5%
9	Do you trust a Robot Surgeon to operate on you?	50%	50%	0%
10	Are you worried that robots will steal your job?	36%	50%	14%
11	Do you think that machines can be biased?	46%	40%	14%
12	Would you send your kids to School in a self-driving bus?	9%	77%	14%
13	Would you like a robot to support your health care in a hospital?	50%	46%	4%
14	Should a machine be liable in a court of law?	23%	64%	4%
15	Would you like to be interviewed by a robot police officer?	41%	59%	0%
16	Do you think robots can predict cancer better than one human doctor?	50%	23%	27%
17	Would you like to be taught a course by a Robot tutor?	46%	41%	13%
18	Would you prefer a robot pet to a real one?	36%	64%	0%
19	Would you let AI choose an outfit for you for your best friend's wedding?	50%	41%	9%
20	Would using facial recognition technology at your child's school make you feel your child was safer in school?	73%	23%	4%

D. Discussion

One interesting observation was in the results to question 6, "Would you trust a self-driving car?", where 27% people said yes and 73% abstained. In discussion, there was more awareness of self-driving cars through media coverage than other topic areas. There was no clear indicator of trust in technologies that supported service industries over those which would have a direct impact on the life/ death of an individual. People stated they would rather trust a pizza making or pizza delivery robot, picking out wedding clothes

or ones that provided an administrative role (Eldercare or receptionist) or relieved humans from being placed in a high risk scenario i.e. space travel. Surprisingly, 46% of people trusted drones to spy on people and 64% thought machines should not be liable in a court of law. 46% of people did believe that machines could be biased, 40% did not and 14% abstained which gave a mixed set of opinions. The personal element provided by humans was reflected in the answers to question 18, where 64% of people preferred a real pet over a robot imitation and 46% would not like to be taught by a human tutor (question 17).

During the debate, it was noted that some members of the public wrote comments on their question sheets. For example, in Reponses to question 10, "Are you worried that robots will steal your job?", two independent people wrote "they already have" and "I was made redundant in manufacturing last year to a machine. I was no longer required – lost self-esteem". During the debate, common questions raised were "What is Computational Intelligence / Artificial Intelligence?; How does it learn / work?; Is it really better than a human?" – providing a clear indicator that further education sessions would be useful. It is noted that this is a small sample, and not a well-defined scientific study, but what it does show is that there is need to provide a platform for education / training for the general public "in the street" about both the strengths and limitations of CI technologies and innovations, so they can make informed ethical choices. Figure 1 shows the public distribution of answers across all 20 questions, showing generally across all topics.

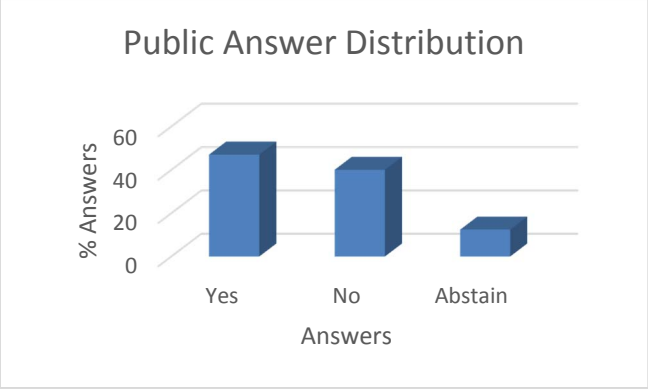


Fig. 1. Public Answer Summary

IV. Towards Trustworthy CI Systems

Developing standards and ethical processes and procedures that deal with trust is and will be a very topical research for the foreseeable future. Tschopp's perception is that trust can be developed through being able to measure the capabilities of AI and highlights a number of current projects such as A-IQ which is used to deduce an AI's level of competence over time [1]. Finkel, chief Scientist of Australia, believes the introduction of an AI Trustmark (inspired by Alan Turing) would indicate that both the vendor and the product were trustworthy in that they complied with standards in the field [42]. Other initiatives include ***morse.ai***, one of a new breed of ethical design led artificial intelligence companies whose aim is promote the use of ethical AI solutions within companies [43] through the development of a morse.code – a *"framework that empowers people and retains control of decisions in human hands"* [43].

The following is a list of common areas of current research activities that are trying to build bridges of trust between humans and CI systems:

Standards - A number of working groups have been formed within the IEEE to pursue ethics standards in autonomous systems. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [14] was publically launched in April 2016. Within the CI society, a Task force on the ethics and social implications of CI was setup in mid-2017 which the authors are members of. In conjunction with the IEEE Society on Social Implications of Technology, the global initiative is developing the IEEE P7000™ series of standards to allow for an elevated level of trust between people and technology. The principal goal of the standards is to prioritize ethical concerns and human wellbeing in the development of standards that address critical aspects of autonomous and intelligent technologies. The standards span issues such as data privacy, algorithmic bias, fail-safe design of autonomous systems and Prioritizing Human Well-being in the Age of Artificial Intelligence. There are currently about 14 standards approved or under development [14].

Training and testing datasets to be representative of the domain - The Dataset Nutrition Label Project [44, 45] is run by the MIT Media Lab and the Berkman Klein Center whose focus is to improve data quality by introducing a system that can develop a nutrition label that can be used to measure the fitness of the dataset in terms of suitability to build a model. The label is determined using both qualitative and quantitative data and a first prototype has been developed [45]. Gebru, T et al. [46] has also suggested that standard be developed called "Datasheets for Datasets" which follows a series of in-depth questions providing information about the dataset creation, composition, pre-processing, distribution and ethical and legal implications.

Dealing with Bias - There is much work to be done in dealing with bias as greater understanding about the implications of a biased algorithm are understood and especially highlighted by the media. For example, in 2018, Buolamwini and Gerbu [47] specifically highlighted the issue of gender bias in computer vision systems, through the evaluation of three commercial systems used for gender classification using a new Parliaments Pilot Benchmark dataset. The work found that the commercial classifiers performed poorly with darker females. On the back of this study, IBM launched the AI Fairness 360 (AIF360) - an open-source toolkit comprising of metrics that check *"for unwanted bias in datasets and*

machine learning models." [47]. In August 2018, IBM researchers have also published the idea of a supplier's declaration of conformity (SDoC) for AI services in order to help increase trust with AI systems. They propose that the SDoC is currently voluntary for businesses much like the datasheets for datasets. Other technological superpowers also have similar projects in an effort to convince the public that they can trust them.

Training and education of humans – Human Stakeholders include: 1) the operators of the CI system, whom require training to help them to understand the implications of bias when interpreting the results produced and empowering them to be able to understand the how the decision was made; 2) the general public whom should be given the opportunity to free education, especially on how CI autonomous systems impact their daily lives and their rights; 3) the scientists and academics who train, validate and test CI systems on how to ensure their algorithms are fair, their datasets are not biased within the context the systems will be used autonomously and that machine based decisions can be explained.

V. CONCLUSIONS

The prime objective of this paper was to examine the current state of trust in CI systems, by sharing the results of a small case study where members of the public voluntary debated the types of CI systems they would trust in 2018. Trust in CI is currently being addressed through the development of standards, new algorithms and metrics that measure fairness and bias, data sheets for open and honest reporting of training, validation and testing data and through commercial industry operating voluntary codes of ethics. However, what appears to be missing is the education of the current general public on what CI / AI systems actually are, how do they work and how do they affect their everyday life. To a person who has not studied or spent time with technology, a bridge needs to be built so that they have the same opportunities to ask questions and understand the answers on how CI systems make decisions / suggestions in their everyday life. Future work will involve the design and delivery and evaluation of a series of short workshops which provide the basics in layman's terms of how CI / AI systems work etc. targeting the everyday person on the street regardless of age or educational level. In addition, future research will also incorporate questions in future surveys associated with risk taking to see if the public's perceptions of "trust" and "risk" are synonymous.

ACKNOWLEDGEMENTS

The study in this paper was supported by Manchester Metropolitan University, IEEE Women in Engineering United Kingdom and Ireland and IEEE Women in Computational Intelligence and Manchester Science Museum. Special thanks go to PhD students Joel Dokmegand and Nicholas Tousaint who supported the Great Computational Intelligence debate on the day.

REFERENCES

[1] Tschopp, M. (2018) On Trust in AI A Systemic Approach, [online] Available: https://www.scip.ch/en/?labs.20180823 [Accessed: 16/12/2018]

[2] Bezdek, J.C., 2013. The history, philosophy and development of computational intelligence (How a simple tune became a monster hit). Computational intelligence. Oxford, UK: Eolss Publishers.

[3] Rousseau, D.M., Sitkin, S.B., Burt, R.S., and Camerer, C. Not so different after all: A cross-discipline view of trust, Academy of Management Review 23, 3 (1998), 393-404.

[4] Luhmann, N. (1979). Trust and power. Chichster: Wiley.

[5] Gambetta, D. (1998). In D. Gambetta (Ed.), Can we trust trust? in Trust: making and breaking cooperative relations (pp. 213-238). Oxford: Basil Blackwell.

[6] Hacker, Philipp, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law (April 18, 2018). Common Market Law Review, Forthcoming. Available: https://ssrn.com/abstract=3164973

[7] Corbett-Davis, S. Goel, Sharad, (2018), The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, Adapated from 19th Conference on Economics and Computation (EC 2018) and at the 35th International Conference on Machine Learning (ICML 2018), [online] Available: https://5harad.com/papers/fair-ml.pdf, Accessed: 06/12/2018.

[8] Robinette, P., Howard, A., & Wagner, A. R. (2017). Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed? In Autonomy and Artificial Intelligence: A Threat or Savior? (pp. 129-155). Springer, Cham.

[9] Borenstein, J., Wagner, A., & Howard, A. (2017). A case study in caregiver overtrust of paediatric healthcare robots. In RSS Workshop on Morality and Social Trust in Autonomous Robots.

[10] Niesche, C. Robots learning bedside manner and it's personal, (2018), [online] Available: https://www.afr.com/news/special-reports/robots-learning-bedside-manner-and-its-personal-20181123-h189ec [Accessed: 13/12/2018]

[11] Seeger, A. M., & Heinzl, A. (2018). Human versus Machine: Contingency Factors of Anthropomorphism as a Trust-Inducing Design Strategy for Conversational Agents. In Information Systems and Neuroscience (pp. 129-139). Springer, Cham.

[12] Yearsley, L., "We Need to Talk About the Power of AI to Manipulate Humans", MIT Technology Review, 5 June 2017

[13] Bhakta, R., Savin-Baden, M., & Tombs, G. (2014, June). Sharing Secrets with Robots? In EdMedia: World Conference on Educational Media and Technology (pp. 2295-2301). Association for the Advancement of Computing in Education (AACE).

[14] Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2018) [online], Available: https://ethicsinaction.ieee.org/, [Accessed 29/12/2018].

[15] Northpointe Institute for Public Management (2012). COMPAS Risk & Need Assessment System: Selected Questions Posed by Inquiring Agencies. Traverse City, MI: Northpointe.

[16] STATE v. LOOMIS, (2016), [online] Available: https://caselaw.findlaw.com/wi-supreme-court/1742124.html, [Accessed 30/12/2018]

[17] Liu, M. (2017), Supreme Court refuses to hear Wisconsin predictive crime assessment case [online], Available: https://www.jsonline.com/story/news/crime/2017/06/26/supreme-court-refuses-hear-wisconsin-predictive-crime-assessment-case/428240001/ [Accessed 30/12/2018]

[18] Ryosuke Hanada, (2016), Is it OK to abuse, trust or make love to a robot?, Nikkei Asian Review, [online], Available:

https://asia.nikkei.com/Business/Technology/Is-it-OK-to-abuse-trust-or-make-love-to-a-robot [Accessed 30/12/2018]

[19] Harrison, S. (2017), Can we trust legal advice from a Robot, Lawyer Monthly, June 30, 2017, [online] Available: https://www.lawyer-monthly.com/2017/06/can-we-trust-legal-advice-from-a-robot/ [Accessed 30/12/2018]

[20] Pagallo, U. (2010), Robotrust and Legal Responsibility, Know Techn Pol (2010) 23:367-379.

[21] Graziadei, M. (1996) cited in Ugo Pagallo, Robotrust and Legal Responsibility, Know Techn Pol (2010) 23:367-379.

[22] Self Driving Wheel Chairs (2018) [online] Available at Source: https://spectrum.ieee.org/the-human-os/biomedical/devices/selfdriving-wheelchairs-debut-in-hospitals-and-airports, [Accessed: 20/11/2018].

[23] Tobe, F. Where Are the Elder Care Robots?, [online] Available at: https://spectrum.ieee.org/automaton/robotics/home-robots/where-are-the-eldercare-robots, [Accessed: 23/11/2018].

[24] Galser, A. Robots will start delivering food to doorsteps in Silicon Valley and Washington, D.C., today Available: Source: https://www.recode.net/2017/1/18/14306674/starship-robot-food-delivery-washington-dc-silicon-valley, Accessed: 23/11/2018].

[25] Guardian, Ban on Killer Robots urgently needed say scientists. [online] Available: https://www.theguardian.com/science/2017/nov/13/ban-on-killer-robots-urgently-needed-say-scientists, Accessed: 23/11/2018].

[26] Guardian, Robots 'could replace 250,000 UK public sector workers' [online], Available: https://www.theguardian.com/technology/2017/feb/06/robots-could-replace-250000-uk-public-sector-workers, [Accessed: 25/11/2018].

[27] Mercedes-Benz , The Mercedes-Benz F 015 Luxury in Motion [online] Available: https://www.mercedes-benz.com/en/mercedes-benz/innovation/research-vehicle-f-015-luxury-in-motion/, [Accessed: 25/11/2018].

[28] Statt, N. Zume's robot pizzeria could be the future of workplace automation, [online] Available: https://www.theverge.com/2017/6/28/15882852/zume-pizza-doughboy-robot-automation-future-food-delivery, [Accessed: 25/11/2018].

[29] Ackerman, E, JAXA Wants Telepresence Robots for In-Space Construction and Exploration (2018), [online] Available: https://spectrum.ieee.org/automaton/robotics/space-robots/jaxa-wants-telepresence-robots-for-inspace-construction-and-exploration, [Accessed: 20/11/2018].

[30] Tomorrows World, Could robotics revolutionise the NHS? (2018), [online] Available: http://www.bbc.co.uk/guides/zssrsg8, [Accessed: 25/11/2018].

[31] Jenkins, S. (2018), Worrying about robots stealing our jobs? [online], Available: https://www.theguardian.com/commentisfree/2018/aug/20/robots-stealing-jobs-digital-age, [Accessed: 25/11/2018].

[32] Turalt, AI isn't biased. We are!, (2018), [online] Available: https://medium.com/@turalt/ai-isnt-biased-we-are-b74ec94d1698, [Accessed: 25/11/2018].

[33] Schwab, K. Would you send your kids to School in a self-driving bus? (2017), [online] Available: https://www.fastcompany.com/90150756/would-you-send-your-kids-to-school-on-a-self-driving-school-bus, [Accessed: 25/11/2018].

[34] Ackerman, E. Moxi Prototype from Diligent Robotics Starts Helping Out in Hospitals, (2018), [online] Available: https://spectrum.ieee.org/automaton/robotics/industrial-robots/moxi-prototype-from-diligent-robotics-starts-helping-out-in-hospitals, Accessed: 25/11/2018].

[35] The Economist, AI am the law, (2015), [online] Available: https://www.economist.com/technology-quarterly/2005/03/10/ai-am-the-law, [Accessed: 25/11/2018].

[36] BBC, Robot police officer goes on duty in Dubai, (2017), [online] Available: https://www.bbc.co.uk/news/technology-40026940, [Accessed: 25/11/2018].

[37] Donnelly, L., Robots are better than doctors at diagnosing some cancers, major study finds, (2018), [online] Available: https://www.telegraph.co.uk/news/2018/05/28/robots-better-doctors-diagnosing-cancers-major-study-finds/ [Accessed: 25/11/2018].

[38] Daily Post, Say good morning to Monica robot maa'm, kids, (2016), [online] Available: https://dailypost.in/news/lifestyle/say-good-morning-monica-robot-maam-kids/, [Accessed: 25/11/2018].

[39] Dador, D. Robotic pets helps seniors reduce stress and feelings of loneliness, (2018), [online] Available: https://abc7.com/health/robotic-pets-helps-seniors-reduce-stress-and-feelings-of-loneliness/4363293/ [Accessed: 25/11/2018].

[40] Cartner-Morley, J. Do robots dream of Prada? How artificial intelligence is reprogramming fashion, (2018), [online] Available: https://www.theguardian.com/fashion/2018/sep/15/do-robots-dream-of-prada-how-artificial-intelligence-is-reprogramming-fashion, Accessed: 25/11/2018]. Image featured in the article is by Alan Davidson/Rex/Shutterstock.

[41] Lerman, R. (Seattle Times), As facial-recognition technology grows, so does wariness about privacy. Use at a school in Seattle fuels debate, (2018), [online] Available: https://www.seattletimes.com/business/technology/as-facial-recognition-technology-grows-so-does-wariness-about-privacy/ [Accessed: 25/11/2018].

[42] Finkel, A. (2018) What will it take for us to trust AI?, [online] Available: https://www.weforum.org/agenda/2018/05/alan-finkel-turing-certificate-ai-trust-robot [Accessed: 16/12/2018.]

[43] Ryan, M (2019), Morse.ai, [online] Available: http://www.morse.ai/#page03 [Accessed: 16/12/2018]

[44] Holland, S. Hosny, A. Newman,S. Joseph, J. Chmielinski, K. (2018), The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards, [online], Available: https://arxiv.org/abs/1805.03677, [Accessed: 06/12/2018].

[45] The Dataset Nutrition Label Project, [online], Available: http://datanutrition.media.mit.edu/, [Accessed: 06/12/2018].

[46] Gebru1, T. Morgenstern, J. Vecchione, B. WortmanVaughan, J. Wallach, H. Daumé, H. Crawford, K. (2018), Datasheets for Datasets, [online], Available: https://arxiv.org/pdf/1803.09010.pdf, [Accessed: 06/12/2018].

[47] Buolamwini, J. Gebru, T. (2018), Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Conference on Fairness, Accountability, and Transparency Proceedings of Machine Learning Research 81, pp, 1 – 15.

[48] Varshney, K. (2018), Introducing AI Fairness 360, [online] Available: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/ [Accessed: 16/12/2018]

[49] Hind,M. Mehta,S. Mojsilovi´c, A. Nair, R. Ramamurthy,K. Olteanu, A. Varshney. K. (2018), Increasing Trust in AI Services through Supplier's Declarations of Conformity [online], Available: https://arxiv.org/abs/1808.07261, [Accessed: 30/12/2018]